(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 756 006 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
29.01.1997 Bulletin 1997/05

(21) Application number: **96109204.6**

(22) Date of filing: **07.06.1996**

(51) Int. Cl.[6]: **C12N 15/31**, C12N 15/10, C12N 5/10, C12N 15/85, C12P 21/08, C07K 14/30, C07K 16/12, C12Q 1/68, A61K 39/395

(84) Designated Contracting States:
AT BE CH DE DK ES FI FR GB GR IE IT LI LU MC NL PT SE

(30) Priority: **07.06.1995 US 488018**
07.06.1995 US 473545
19.10.1995 US 545528

(71) Applicants:
- **THE INSTITUTE FOR GENOMIC RESEARCH**
  Rockville, Maryland 20850 (US)
- **THE JOHNS HOPKINS UNIVERSITY**
  Baltimore, MD 21205 (US)
- **THE UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL**
  Chapel Hill, North Carolina 27599-4105 (US)

(72) Inventors:
- **Fraser, Claire M.**
  Rockville, Maryland 20850 (US)
- **Adams, Mark D.**
  N. Potomac, Maryland (US)
- **Gocayne, Jeannine D.**
  Silver Springs, Maryland 20902 (US)
- **Hutchison, Clyde A., III**
  Chapel Hill, North Carolina 27514 (US)
- **Smith, Hamilton O.**
  Towson, Maryland 21204 (US)
- **Venter, J. Craig**
  Rockville, Maryland 20850 (US)
- **White, Owen**
  Gaithersburg, Maryland 20878 (US)

(74) Representative: **VOSSIUS & PARTNER**
Siebertstrasse 4
81675 München (DE)

(54) Nucleotide sequence of the mycoplasma genitalium genome, fragments thereof, and uses thereof

(57) The present invention provides the nucleotide sequence of the entire genome of *Mycoplasma genitalium*, SEQ ID NO:1. The present invention further provides the sequence information stored on computer readable media, and computer-based systems and methods which facilitate its use. In addition to the entire genomic sequence, the present invention identifies protein encoding fragments of the genome, and identifies, by position relative to two (2) genes known to flank the origin of replication, any regulatory elements which modulate the expression of the protein encoding fragments of the *Mycoplasma genitalium* genome.

EP 0 756 006 A2

Printed by Rank Xerox (UK) Business Services
2.13.13/3.4

The nucleotide sequence information provided in SEQ ID NO:1 was obtained by sequencing the *Mycoplasma genitalium* genome using a megabase shotgun sequencing method. The nucleotide sequence provided in SEQ ID NO:1 is a highly accurate, although not necessarily a 100% perfect, representation of the nucleotide sequence of the *Mycoplasma genitalium* genome.

As discussed in detail below, using the information provided in SEQ ID NO:1 and in Tables 1(a), 1(c) and 2 together with routine cloning and sequencing methods, one of ordinary skill in the art would be able to clone and sequence all "representative fragments" of interest including open reading frames (ORFs) encoding a large variety of *Mycoplasma genitalium* proteins. In very rare instances, this may reveal a nucleotide sequence error present in the nucleotide sequence disclosed in SEQ ID NO:1. Thus, once the present invention is made available (i.e., once the information in SEQ ID NO:1 and Tables 1(a), 1(c) and 2 have been made available), resolving a rare sequencing error in SEQ ID NO:1 would be well within the skill of the art. Nucleotide sequence editing software is publicly available. For example, Applied Biosystem's (AB) AutoAssembler™ can be used as an aid during visual inspection of nucleotide sequences.

Even if all of the very rare sequencing errors in SEQ ID NO:1 were corrected, the resulting nucleotide sequence would still be at least 99.9% identical to the nucleotide sequence in SEQ ID NO:1.

The nucleotide sequences of the genomes from different strains of *Mycoplasma genitalium* differ slightly. However, the nucleotide sequence of the genomes of all *Mycoplasma genitalium* strains will be at least 99.9% identical to the nucleotide sequence provided in SEQ ID NO:1.

Thus, the present invention further provides nucleotide sequences which are at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 in a form which can be readily used, analyzed and interpreted by the skilled artisan. Methods for determining whether a nucleotide sequence is at least 99.9% identical to the nucleotide sequence of SEQ ID NO:1 are routine and readily available to the skilled artisan. For example, the well known fasta algothrithm (Pearson and Lipman, *Proc. Natl. Acad. Sci. USA 85*:2444 (1988)) can be used to generate the percent identity of nucleotide sequences.

Computer Related Embodiments

The nucleotide sequence provided in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1 may be "provided" in a variety of mediums to facilitate use thereof. As used herein, provided refers to a manufacture, other than an isolated nucleic acid molecule, which contains a nucleotide sequence of the present invention, i.e., the nucleotide sequence provided in SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1. Such a manufacture provides the *Mycoplasma genitalium* genome or a subset thereof (e.g., a *Mycoplasma genitalium* open reading frame (ORF)) in a form which allows a skilled artisan to examine the manufacture using means not directly applicable to examining the *Mycoplasma genitalium* genome or a subset thereof as it exists in nature or in purified form.

In one application of this embodiment, a nucleotide sequence of the present invention can be recorded on computer readable media. As used herein, "computer readable media" refers to any medium which can be read and accessed directly by a computer. Such media include, but are not limited to: magnetic storage media, such as floppy discs, hard disc storage medium, and magnetic tape; optical storage media such as CD-ROM; electrical storage media such as RAM and ROM; and hybrids of these categories such as magnetic/optical storage media A skilled artisan can readily appreciate how any of the presently known computer readable mediums can be used to create a manufacture comprising computer readable medium having recorded thereon a nucleotide sequence of the present invention.

As used herein, "recorded" refers to a process for storing information on computer readable medium. A skilled artisan can readily adopt any of the presently know methods for recording information on computer readable medium to generate manufactures comprising the nucleotide sequence information of the present invention.

A variety of data storage structures are available to a skilled artisan for creating a computer readable medium having recorded thereon a nucleotide sequence of the present invention. The choice of the data storage structure will generally be based on the means chosen to access the stored information. In addition, a variety of data processor programs and formats can be used to store the nucleotide sequence information of the present invention on computer readable medium. The sequence information can be represented in a word processing text file, formatted in commercially-available software such as WordPerfect and MicroSoft Word, or represented in the form of an ASCII file, stored in a database application, such as DB2, Sybase, Oracle, or the like. A skilled artisan can readily adapt any number of dataprocessor structuring formats (e.g. text file or database) in order to obtain computer readable medium having recorded thereon the nucleotide sequence information of the present invention.

By providing the nucleotide sequence of SEQ ID NO:1, a representative fragment thereof, or a nucleotide sequence at least 99.9% identical to SEQ ID NO:1 in computer readable form, a skilled artisan can routinely access the sequence information for a variety of purposes. Computer software is publicly available which allows a skilled artisan to access sequence information provided in a computer readable medium. The examples which follow demonstrate how software which implements the BLAST (Altschul *et al., J. Mol. Biol. 215*:403-410 (1990)) and BLAZE (Brutlag *et al., Comp. Chem. 17*:203-207 (1993)) search algorithms on a Sybase system was used to identify open reading frames

(ORFs) within the *Mycoplasma genitalium* genome which contain homology to ORFs or proteins from other organisms. Such ORFs are protein encoding fragments within the *Mycoplasma genitalium* genome and are useful in producing commercially important proteins such as enzymes used in fermentation reactions and in the production of commercially useful metabolites.

The present invention further provides systems, particularly computer-based systems, which contain the sequence information described herein. Such systems are designed to identify commercially important fragments of the *Mycoplasma genitalium* genome.

As used herein, "a computer-based system" refers to the hardware means, software means, and data storage means used to analyze the nucleotide sequence information of the present invention. The minimum hardware means of the computer-based systems of the present invention comprises a central processing unit (CPU), input means, output means, and data storage means. A skilled artisan can readily appreciate that any one of the currently available computer-based system are suitable for use in the present invention.

As stated above, the computer-based systems of the present invention comprise a data storage means having stored therein a nucleotide sequence of the present invention and the necessary hardware means and software means for supporting and implementing a search means. As used herein, "data storage means" refers to memory which can store nucleotide sequence information of the present invention, or a memory access means which can access manufactures having recorded thereon the nucleotide sequence information of the present invention.

As used herein, "search means" refers to one or more programs which are implemented on the computer-based system to compare a target sequence or target structural motif with the sequence information stored within the data storage means. Search means are used to identify fragments or regions of the *Mycoplasma genitalium* genome which match a particular target sequence or target motif. A variety of known algorithms are disclosed publicly and a variety of commercially available software for conducting search means are available and can be used in the computer-based systems of the present invention. Examples of such software includes, but is not limited to, MacPattern (EMBL), BLASTN and BLASTX (NCBIA). A skilled artisan can readily recognize that any one of the available algorithms or implementing software packages for conducting homology searches can be adapted for use in the present computer-based systems.

As used herein, a "target sequence" can be any DNA or amino acid sequence of six or more nucleotides or two or more amino acids. A skilled artisan can readily recognize that the longer a target sequence is, the less likely a target sequence will be present as a random occurrence in the database. The most preferred sequence length of a target sequence is from about 10 to 100 amino acids or from about 30 to 300 nucleotide residues. However, it is well recognized that searches for commercially important fragments of the *Mycoplasma genitalium* genome, such as sequence fragments involved in gene expression and protein processing, may be of shorter length.

As used herein, "a target structural motif," or "target motif," refers to any rationally selected sequence or combination of sequences in which the sequence(s) are chosen based on a three-dimensional configuration which is formed upon the folding of the target motif. There are a variety of target motifs known in the art. Protein target motifs include, but are not limited to, enzymic active sites and signal sequences. Nucleic acid target motifs include, but are not limited to, promoter sequences, hairpin structures and inducible expression elements (protein binding sequences).

A variety of structural formats for the input and output means can be used to input and output the information in the computer-based systems of the present invention. A preferred format for an output means ranks fragments of the *Mycoplasma genitalium* genome possessing varying degrees of homology to the target sequence or target motif. Such presentation provides a skilled artisan with a ranking of sequences which contain various amounts of the target sequence or target motif and identifies the degree of homology contained in the identified fragment.

A variety of comparing means can be used to compare a target sequence or target motif with the data storage means to identify sequence fragments of the *Mycoplasma genitalium* genome. In the present examples, implementing software which implement the BLAST and BLAZE algorithms (Altschul *et al., J. Mol. Biol. 215*:403-410 (1990)) was used to identify open reading frames within the *Mycoplasma genitalium* genome. A skilled artisan can readily recognize that any one of the publicly available homology search programs can be used as the search means for the computer-based systems of the present invention.

One application of this embodiment is provided in Figure 2. Figure 2 provides a block diagram of a computer system 102 that can be used to implement the present invention. The computer system 102 includes a processor 106 connected to a bus 104. Also connected to the bus 104 are a main memory 108 (preferably implemented as random access memory, RAM) and a variety of secondary storage devices 110, such as a hard drive 112 and a removable medium storage device 114. The removable medium storage device 114 may represent, for example, a floppy disk drive, a CD-ROM drive, a magnetic tape drive, etc. A removable storage medium 116 (such as a floppy disk, a compact disk, a magnetic tape, etc.) containing control logic and/or data recorded therein may be inserted into the removable medium storage device 114. The computer system 102 includes appropriate software for reading the control logic and/or the data from the removable medium storage device 114 once inserted in the removable medium storage device 114.

A nucleotide sequence of the present invention may be stored in a well known manner in the main memory 108, any of the secondary storage devices 110, and/or a removable storage medium 116. Software for accessing and

processing the genomic sequence (such as search tools, comparing tools, etc.) reside in main memory 108 during execution.

Biochemical Embodiments

Another embodiment of the present invention is directed to isolated fragments of the *Mycoplasma genitalium* genome. The fragments of the *Mycoplasma genitalium* genome of the present invention include, but are not limited to fragments which encode peptides, hereinafter open reading frames (ORFs), fragments which modulate the expression of an operably linked ORF, hereinafter expression modulating fragments (EMFs), fragments which mediate the uptake of a linked DNA fragment into a cell, hereinafter uptake modulating fragments (UMFs), and fragments which can be used to diagnose the presence of *Mycoplasma genitalium* in a sample, hereinafter diagnostic fragments (DFs).

As used herein, an "isolated nucleic acid molecule" or an "isolated fragment of the *Mycoplasma genitalium* genome" refers to a nucleic acid molecule possessing a specific nucleotide sequence which has been subjected to purification means to reduce, from the composition, the number of compounds which are normally associated with the composition. A variety of purification means can be used to generated the isolated fragments of the present invention. These include, but are not limited to methods which separate constituents of a solution based on charge, solubility, or size.

In one embodiment, *Mycoplasma genatalium* DNA can be mechanically sheared to produce fragments of 15-20 kb in length. These fragments can then be used to generate an *Mycoplasma genitalium* library by inserting them into lambda clones as described in the Examples below. Primers flanking, for example, an ORF provided in Table 1(a), 1(c) or 2 can then be generated using nucleotide sequence information provided in SEQ ID NO:1. PCR cloning can then be used to isolate the ORF from the lambda DNA library. PCR cloning is well known in the art. Thus, given the availability of SEQ ID NO:1, Table 1(a), 1(c) and Table 2, it would be routine to isolate any ORF or other representative fragment of the present invention.

The isolated nucleic acid molecules of the present invention include, but are not limited to single stranded and double stranded DNA, and single stranded RNA.

As used herein, an "open reading frame," ORF, means a series of triplets coding for amino acids without any termination codons and is a sequence translatable into protein. Tables 1(a), 1(b), 1(c) and 2 identify ORFs in the *Mycoplasma genitalium* genome. In particular, Table 1(a) indicates the location of ORFs (i.e., the adresses) within the *Mycoplasma genitalium* genome which encode the recited protein based on homology matching with protein sequences from the organism appearing in parentheticals (see the fifth column of Table 1(a)).

The first column of Table 1(a) provides the "UID" (an arbitrary identification number) of a particular ORF. The second and third columns in Table 1(a) indicate an ORFs position in the nucleotide sequence provided in SEQ ID NO:1. One of ordinary skill in the art will recognize that ORFs may be oriented in opposite directions in the *Mycoplasma genitalium* genome. This is reflected in columns 2 and 3.

The fourth column of Table 1(a) provides the accession number of the database match for the ORF. As indicated above, the fifth column of Table 1(a) provides the name of the database match for the ORF.

The sixth column of Table 1(a) indicates the percent identity of the protein encoded for by an ORF to the corresponding protein from the orgaism appearing in parentheticals in the fifth column. The seventh column of Table 1(a) indicates the percent similarity of the protein encoded for by an ORF to the corresponding protein from the organism appearing in parentheticals in the fifth column. The concepts of percent identity and percent similarity of two polypeptide sequences are well understood in the art. For example, two polypeptides 10 amino acids in length which differ at three amino acid positions (e.g., at positions 1,3 and 5) are said to have a percent identity of 70%. However, the same two polypeptides would be deemed to have a percent similarity of 80% if, for example at position 5, the amino acids moieties, although not identical, were "similar" (i.e., possessed similar biochemical characteristics). The eighth column in Table 1(a) indicates the length of the ORF in nucleotides.

Table 1(b) is a list of ORFs that have database matches to previously published *Mycoplasma genitalium* sequences over the full length of the ORF. The talbe headings for Table 1(b) are identical for Table 1(a) with the following two exceptions: (II) The heading for the eighth column in Table 1(a) (i.e., nucleotide length of the ORF) has been replaced with the following in Table 1(b): "Match_info". "Match_info" refers to the coordinates of the match of the ORF and the previously published *Mycoplama genitalium* sequence. For example, "MG002 (1-930 of 930) GB:U09251 (298-1227 of 6140)," indicates that for ORF MG002, which is 930 nucleotides in length, there is a database match to accession number GB:U09251, which has a total length of 6140 nucleotides. The ORF matches this accession from position 298 to 1227. (II) Where an ORF shows homology matches for both a previously published *Mycoplasma genitalium* sequence and a previously published sequence from a different organism, columns 3, 4, 5, and 6 of Table 1(b) respectively provide the accession number, protein name (and organism in parentheticals), percent identity and percent similarity for the "other organism," rather than for the previously published *Mycoplasma genitalium* sequence. (However, in this scenario, the accession number for the *Mycoplasma genitalium* sequence is still provided in column 8.)

Table 1(c) provides ORFs having database matches to previously published *Mycoplasma genitalium* sequences

```
ACCATTATTA TGATATTGAA AATTTTGTTC CTCTTGAAAA TATCTCTCTT TTTTTGGTTT  578880
TCCAGAAAAA TTTGATGAAA AAGATTTTCC TTCATTTCAA TTTTCAAGAT TATTTTCATT  578940
TTGTTGATTT ATTTGCTCAG GCTGTTGAAA TGAATTATTT TTTGATCAAA AAGATTTTGG  579000
AAAGGTTTTT TCAAAAGCAG ATAAAGGTCC AAAATCAAAT GAAGATGAAT CTTTGTCAAA  579060
AGATGTTTCT TCTCTTTTTG ACAAATTTTG TTTTTGATTA AACTTATTTT TATTTTGGGG  579120
TGTTACTTTT TCTTTTATGG AAAACAAATC TTCTTCTAAA AGACTTTGTT CTGGGTCATC  579180
ATCTTGTGCT AAATCAAAGA AAAAACGTTT CTTTTTGTTA TTAATGGACA TTGTAATTTG  579240
CTAAATTTAG GATTTCTTTT GTTATTTCTA AATACTCATT TAGATATTTT TTACTTGGTG  579300
ATGATACTAA TGATATTGGC AATTTTTCAT AACCTACAGC TGCTGATGAT TTTGATGTCA  579360
GAGAAACAAA ATTTTTAGAA AAAGCTACAT TATTTTTTTT AGCTTTTGTT TTAGCTAAAT  579420
CTATTACTTC ATTATGAAGA CGAGTACGAA CGTTAACTTT TGTAGGAACT AAAATAGTTT  579480
TAAGATTTGT ATTTTTTTCC TTAAATGTAT CTATTGTTTC AACTATTCTC ATCAAACCTA  579540
GCATCGAATA TTGATCTGGT TCAAAGGGAA TAACTATGAC ATCTGATAAA CTCATTGCAG  579600
TAGAAACTAA AGTTGCCATA TTTGGTGGTG TATCTAATAA AACAAATTCA TATCTTTTTG  579660
CTAGTTGCTT AACTATTTCT GCTATATCTG AGGCCTTATA TTTTTTACGT GATATGTCTA  579720
TATCAGCAAA ATTAAGTTCA AAATTACAAG GAAGAATATC AAGTCCCTCA TATACAGATA  579780
GCAAGCAATC ATCTATTTCA ATGAAATTAT TTGAACCACT GAATTTTGGA ACCTTCAACA  579840
AAATGTCAAT TAACGTGTTA TTCAATCTTT CAGGGTTTTG TCCAAATGAT GCAGAAACAT  579900
TCCCCTGCCC GTCAAGATCA AGAATGACTT TTCGCCTTTC TGGACAAAGT TTAACCAATG  579960
ATCCTGCAAC ATTAGTTGCC ATTGTAGTTT TTAATACGCC GCCTTTATTA TTTACAAAAG  580020
AAATGATCAT ATATTTAAAT GATTATAATA TTTCTTTAAT ACTAAAAAAA TAC         580073
```

Claims

1. Computer readable medium having recorded thereon a nucleic acid sequence selected from the group consisting of:

   (a) the nucleic acid sequence depicted in SEQ ID NO:1;
   (b) a representative fragment of the nucleic acid sequence of (a);
   (c) a nucleic acid sequence at least 99.9 % identical to the nucleic acid sequence of (a);
   (d) a DNA molecule of (b) as depicted in Tables 1a, 1c and 2; and
   (e) a degenerate variant of (d).

2. The computer readable medium of claim 1, wherein said medium is selected from the group consisting of:

   (a) a floppy disc;
   (b) a hard disc;
   (c) random access memory (RAM);
   (d) read only memory (ROM); and
   (e) CD-ROM.

3. A computer-based system for identifying fragments of the *Mycoplasma* genome of commercial importance comprising the following elements:

(a) a data storage means comprising a nucleic acid sequence as described in claim 1;
(b) search means for comparing a target sequence to the nucleotide sequence of the date storage means of step (a) to identify homologous sequence (s); and
(c) retrieval means for obtaining said homologous sequence (s) of step (b).

4. A method for identifying commercially important nucleic acid fragments of the *Mycoplasma* genome comprising the step of comparing a database comprising a nucleotide sequence as described in claim 1 with a target sequence to obtain a nucleic acid molecule comprised of a complementary nucleotide sequence to said target sequence, wherein said target sequence is not randomly selected.

5. A method for identifying an expression modulating fragment of the *Mycoplasma* genome comprising the step of comparing a database comprising the nucleotide sequence as described in claim 1 with a target sequence to obtain a nucleic acid molecule comprised of a complementary nucleotide sequence to said target sequence, wherein said target sequence comprises sequences known to regulate gene expression.

6. A protein-encoding nucleic acid fragment of the *Mycoplasma genitalium* genome, wherein said fragment consists of the nucleic acid sequence referred to in claim 1.

7. A fragment of the *Mycoplasma genitalium* genome, wherein

(a) said fragment modulates the expression of an operably linked open reading frame;
(b) said fragment consists of a nucleotide sequence from about 10 to 200 bases in length which is 5' to any one of the open reading frames depicted in Tables 1a, 1c and 2; or
(c) said fragment consists of a degenerate variant of (b).

8. A nucleic acid molecule encoding a homolog of any one of the fragments of the *Mycoplasma* genome depicted in Tables 1a, 1c and 2, wherein said nucleic acid molecule is produced by the steps of:

(a) screening a genomic library using any one of the fragments of the *Mycoplasma* genome depicted in Tables 1a, 1c and 2 as a target sequence;
(b) identifying members of said library which contain sequences which hybridize to said target sequence;
(c) isolating the nucleic acid molecules from said members identified in step (b).

9. A DNA molecule encoding a homolog of any one of the fragments of the *Mycoplasma* genome depicted in Tables 1a, 1c and 2, wherein said nucleic acid molecule is produced by the steps of:

(a) isolating mRNA, DNA or cDNA produced from an organism;
(b) amplifying nucleic acid molecules whose nucleotide sequence is homologous to amplification primers derived from said fragment of said *Mycoplasma* genome to prime said amplification;
(c) isolating said amplified sequences produced in step (b).

10. A vector comprising any one of the fragments of the *Mycoplasma genitalium* genome selected from the group consisting of the nucleic acid fragments of any one of claims 6 to 9.

11. An organism which has been altered to contain any one of the fragments of the *Mycoplasma* genome of claims 6 to 9.

12. A method for regulating the expression of a nucleic acid molecule comprising the step of covalently attaching 5' to said nucleic acid molecule a nucleic acid molecule consisting of the nucleotide sequence from about 10 to 100 bases 5' to any one of the fragments of the *Mycoplasma* genome depicted in Tables 1a, 1c or 2, or a degenerate variant thereof.

13. A polypeptide encoded by any one of the nucleic acid molecules of claim 6, 8 or 9.

14. An antibody which selectively binds to a polypeptide of claim 13.

15. The antibody of claim 14 which is monoclonal.

16. A method for producing a polypeptide in a host cell comprising the steps of:

(a) incubating a host containing a heterologous nucleic acid molecule whose nucleotide sequence consists of any one of the fragments of the *Mycoplasma genitalium* genome depicted in Table 1a or a degenerate variant thereof, excluding the fragments of SEQ ID NO:1 depicted in Table 1b under conditions where said heterologous nucleic acid molecule is expressed to produce said protein, and
(b) isolating said protein.

17. A pharmaceutical composition comprising the antibody of claim 14 or 15.